DOCUMENT RESUME

ED 248 269 TM 840 575

AUTHOR Marco, Gary L.
TITLE An Evaluation of an Indirect Method of Transforming Item Parameter Estimates from Item Response Theory to a Common Scale.
PUB DATE Apr 84
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Equated Scores; *Estimation (Mathematics); *Latent Trait Theory; Mathematical Models; *Scaling; Test Construction; Test Items; Transformations (Mathematics)
IDENTIFIERS BICAL Computer Program; *Item Parameters; LOGIST Computer Program; Quantile Computer Program; Scholastic Aptitude Test

ABSTRACT

Using raw-to-scaled-score conversions derived from test-score equating to link item-parameter estimates from the one-parameter (Rasch) and three-parameter logistic models, this study evaluated an indirect method for converting item response theory estimates to a common scale. Data were taken from Petersen's Scholastic Aptitude Test (SAT) scale stability study, which involved the chain equating of a test to itself through five intermediary forms. Sample sizes ranged from 2,527 to 2,879. Six SAT-verbal and six SAT-mathematical test forms were used. To estimate item parameters from individual data, the BICAL computer program was used for the one-parameter logistic model and the LOGIST program for the three-parameter model. The Quantile program estimated item parameters from grouped data for the three-parameter model. IRT true score equating provided a criterion to assess the practical consequences of using the three item calibration methods in conjunction with the score-conversion method of item-parameter transformation. Within-model and between-model comparisons showed close agreement with initial methods. While results were promising, further study of item parameter transformations based on score conversion information is needed before operational application in testing programs. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED248269

An Evaluation of an Indirect Method of Transforming Item Parameter Estimates from Item Response Theory to a Common Scale

Gary L. Marco
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper Presented at the Annual Meeting
of the American Educational Research
Association, New Orleans

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G.L. Marco

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

April 1984

TM 810 575

An Evaluation of an Indirect Method of Transforming Item Parameter Estimates from Item Response Theory to a Common Scale[1]

Although IRT methods are often superior to traditional methods for equating and other purposes, the costs of converting from traditional to IRT methods can be substantial. The LOGIST computer program (Wingersky, 1983) and other computer programs used to estimate IRT item parameters for the three-parameter logistic test model take a considerable amount of computer time and thus are expensive to run for large data sets. The cost is particularly high when IRT methods are introduced into an existing testing program because the parameters of a large number of items must not only be estimated using a program like LOGIST but also placed on a common scale through complicated common-item linkages.

Previous research suggests that approximate IRT methods might be useful for estimating item parameters (item calibration) and placing item parameter estimates on a common scale (item-parameter transformation). For example, Kee (1979) compared item parameter and ability estimates obtained from LOGIST and two of Urry's programs, ANCILLES and OGIVIA, with the true parameters from which simulated data had been generated. He found that none of the programs were uniformly superior for parameter estimation. However, the cost of using Urry's programs were only 10% to 15% of the cost of using LOGIST. In a study of PSAT/NMSQT pre-equating, Marco (1977) used the approximate method of item parameter transformation investigated in the current study to avoid the considerable expense of calibrating items from a large number of test forms. He used existing score equating results based on traditional linear equating to scale item parameter estimates from separate applications of LOGIST. Marco found that, except at the upper end of the score scale, the pre-equating based on item parameter estimates

transformed with the use of score equating information results agreed reasonably well with the criterion equatings. Lord (1975) had previously used this method to transform item parameter estimates from a broad range of verbal ability tests to a common scale.

While several methods are available for transforming item parameters from item response theory to a common scale, the usual methods require either common examinees (the same examinees take all items) or common items (item sets taken by different examinees have items in common). The method investigated in this study requires neither common examinees nor common items and thus provides a way to link item parameters when these conditions are not met. However, additional information must be available, namely, information on converting total-test raw scores to a common scale. This method is particularly useful in establishing item banks for those testing programs that have a continuous history of form-to-form test-score equating but have few common-item links among current forms.

The purpose of this study was to evaluate an indirect method of transforming item-parameter estimates from item response theory to a common scale. Although the method had been used previously by Lord (1975) and Marco (1977) to "scale" item parameter estimates, the method has never been evaluated formally. The method involves the use of raw-to-scaled-score conversions derived from test-score equating to link item parameter-estimates. In the present study the method was applied to item parameter estimates from the one-parameter (Rasch) and three-parameter logistic models.

Procedures

The tests and examinee samples for this study were those used for a recent Scholastic Aptitude Test (SAT) scale stability study (Petersen, Cook, & Stocking, 1983). That study investigated several methods for equating scores from six SAT-verbal and six SAT-mathematical test forms including several variants of IRT equating. The study involved the chain equating of a test to itself through five intermediary forms. These data can be used to study item calibration and item-parameter transformation procedures, the focus of the current study, as well as equating procedures.

Tests and Test Scores

The tests consisted of six operational and six equating tests for SAT-verbal and SAT-mathematical, respectively, administered between December 1973 and May 1979. The tests were chosen so that the calibration links formed a closed circle. These tests are identified in Figure 1, which shows the chains of six verbal and six mathematical item calibrations that were used in the study. SAT forms are indicated by upper case letters and equating tests by lower case letters. Each SAT-verbal form except Form V4, which was administered prior to the fall of 1974, had 85 items; Form V4 contained 90 items. A given verbal equating test contained 40 items. Each SAT-mathematical form except Form Y3 had 60 items; and each mathematical equating test except Form fn, 25 items. Due to printing error, SAT-mathematical Form Y3 had 59 items, and mathematical equating test fn had 24 items.

Insert Figure 1 about here

Raw scores on the SAT are formula scores based on the number right minus a fraction of the number wrong, where the fraction is 1/(no. of response options - 1). Raw scores for a particular test form are converted to scaled scores on the 200 to 800 College Board scale by applying the mathematical transformation derived through score equating.

Data Used in the Study

The sample consisted of approximately 2,670 cases for each pairing of an SAT form and an equating test shown in Figure 1. The actual sample sizes ranged from 2,527 to 2,879. The samples were randomly selected from examinees taking the SAT at the respective administrations. Figure 2 shows the data sets that were used in the study. Individual records contained item response data appropriate for use in the various computer programs, which required information on right and wrong responses to each test item. Records also contained information on items omitted and items not reached. Table 1 gives the SAT scaled score means and standard deviations for the samples used in the study.

Insert Figure 2 and Table 1 about here

Study Design

The score-conversion method of item-parameter transformation was used for the chains of six SAT-verbal and SAT-mathematical test forms shown in

Figure 1. The item parameters for Form V4 were transformed to the Form V4 scale through five intermediary forms. The data sets used in this process are identified by asterisks in Figure 2. Since the item parameters for the common item were not needed for deriving transformations based on score conversions, only 7 of the 12 data sets had to be used.

Three methods of item calibration, described in the following section, were used in the study to obtain item-parameter estimates for the items in the SAT-verbal and -mathematical forms represented in the study. The BICAL computer program was used to estimate item parameters from individual data for the one-parameter logistic (Rasch) model, the LOGIST computer program was used to estimate item parameters from individual data for the three-parameter logistic model, and the Quantile computer program, a modification of LOGIST, was used to estimate item parameters from grouped data (fifths) for the three-parameter logistic model.

Three criteria were used to evaluate the quality of the item-parameter estimates derived from score-conversion information. First, the final transformations for the SAT-verbal and -mathematical chains and the resulting item parameter estimates were compared with the identify transformation that would have resulted under ideal conditions. In addition, the final Form V4 item parameters were compared with the initial Form V4 item parameters. BICAL final results were compared with the BICAL initial results, the LOGIST final results were compared with the LOGIST initial results, and the Quantile final results were compared with the Quantile initial results. The various final estimates were also compared with the

initial LOGIST estimates. Thus, both within-model and between-model comparisons were made.

Second, for the BICAL and LOGIST item parameter estimates only, the intermediate and final transformations were compared with the intermediate and final transformations derived through common-items. All 12 data sets given in Figure 2 had been used previously (Petersen et al., 1983; Marco, Wingersky, & Douglass, Note 1) for deriving the BICAL and LOGIST common-items item-parameter transformations. The common-item method of obtaining item-parameter transformations is the usual procedure used in operational testing programs to place item parameters on a common scale and thus provides a good criterion against which to evaluate the results of the score-conversion method of obtaining item-parameter transformations. The BICAL common-item item parameter transformation was derived by setting equal the means from the two separate calibrations involving the items common to any two SAT-verbal or -mathematical SAT forms (Wright & Stone, 1979). For BICAL estimates an additive constant provides the appropriate adjustment to the b's. For example, the item parameter estimates for SAT-verbal Form X2 and Equating Test fe were transformed to the scale for SAT-verbal Form V4 and Equating Test fe by subtracting the mean b of equating test fe for the examinees who took Form X2 from the equating test mean b of equating test fe for the examinees who took Form V4. The transformation for placing the item parameter estimates for any particular form on the V4 scale was obtained simply by summing the constants in the chain. In this way, a transformation was obtained for placing the item parameters for Form V4 on the Form V4 scale indirectly.

The method used to obtain the common-item transformations for LOGIST item parameters was the characteristic-curve method (Stocking & Lord, 1983), which minimizes the average squared difference between number right true scores on the items common to a given pair of test forms for a given group of examinees. The LOGIST transformations are linear, but have two parameters rather than one. The same stepwise sequence described for the score-conversion method was applied to derive final transformations and item parameters for Form V4.

A third criterion was used in order to discover the practical consequences of using the score-conversion method instead of some other method to obtain item-parameter transformations. The results of equating Form V4 final scores to Form V4 initial scores using the transformed item parameters derived by the score-conversion method were compared with the equating results using transformed item parameters placed on scale by other methods. (Most of these equating results were already available from the Petersen et al. (1983) study.) These comparisons were among-model comparisons, since all of the equatings results were on the same scale.

Item Parameter Transformation Based on Score-Conversion Information

The item parameter estimates from separate item calibrations must be transformed to a common scale in order for them to be used together. In the present study this was accomplished using operational score-conversion information in the manner described by Marco (1977). The essential steps are as follows:

1. For a given ability level, $\underline{\theta}_i$, for test i, compute the true number right score, $\underline{R}_i$, by the formula $\underline{P}_g$, where $\underline{P}_g$, the probability of

answering item g correctly, is computed from the estimated parameters for item g.

2. Express $\underline{R}_i$ as a true formula score, $\underline{FS}_i$, under the assumption that all items are answered: $\underline{FS} = \underline{R} - (\underline{N}_5 - \underline{R}_5)/4$ for SAT-verbal (five-choice items) and $\underline{FS} = \underline{R} - (\underline{N}_4 - \underline{R}_4)/3 - (\underline{N}_5 - \underline{R}_5)/4$ for SAT-mathematical (four- and five-choice items), where $\underline{N}$ is the number of test items and the subscript indicates whether four-choice (4) or five-choice (5) items are involved.
3. Transform $\underline{FS}_i$ to the College Board scale ($\underline{S}$) using the operational scaling parameters derived previously when the tests were originally equated (see Table 5): $\underline{S} = \underline{A}_i\underline{FS}_i + \underline{B}_i$.
4. Find the true $\underline{FS}$ on test j corresponding to this scaled score:

$$\underline{FS}_j = (\underline{S} - \underline{B}_j)/\underline{A}_j ,$$

where $\underline{A}_j$ and $\underline{B}_j$ are the scaling parameters (see Table 2).
5. Compute the true $\underline{R}_j$ from the formulas in Step 2.
6. Determine the corresponding ability level $\underline{\theta}_j$ for test j by finding the $\underline{\theta}_j$ for which $\underline{P}_h$ equals $\underline{R}_j$ using the item-parameter estimate for test j.
7. Apply steps 1-6 to approximately 60 equally spaced ability levels between -3 and 3.
8. Determine the straight line relating the $\underline{\theta}_j$'s to the $\underline{\theta}_i$'s in the range -1.75 to 1.75 by setting their means and standard deviations equal. The range is restricted to prevent outliers in the tails of the score range from influencing the results. This process results

is a transformation $\underline{\theta}_i = \underline{A}\underline{\theta}_j + \underline{B}$. Here $\underline{A} = SD(\underline{\theta}_i)/SD(\underline{\theta}_j)$ and $\underline{B} = M(\underline{\theta}_i) - M(\underline{\theta}_j)$, where M and SD stand for mean and standard deviation, respectively.

9. Determine the transformation for placing the item-parameter estimates on test j onto the scale for the item-parameter estimates on test i as follows:

$$\underline{b}_i = \underline{A}\underline{b}_j + \underline{B} \text{ and } \underline{a}_i = \underline{a}_j/\underline{A} .$$

The $\underline{c}$'s are unaffected by scale transformations.

Insert Table 2 about here

The first step in applying the score-conversion method was to transform the item parameters on Form X2 to the score scale established for the item parameters on Form V4. Then the item parameters for Form Y3 were transformed to Form V4 item-parameter scale using the transformed item parameters on Form X2 resulting from the previous step. Continuation of this sequential process resulted finally in transforming the Form V4 item parameters, estimated on a new sample, to the Form V4 item-parameter scale using the transformed item parameters on Form Z5. These transformed item parameters on Form V4 could then be compared with the original Form V4 item parameters; ideally, the item parameters for a given item would be identical. Any discrepancy could be attributed to the particular item calibration and item-parameter transformation methods used in the study.

Item Calibration Methods

There are two separate but related steps required to produce item parameter estimates from different tests on a single scale. The first step, item calibration, is to estimate item parameters. The second step, item parameter transformation - required when item parameters are estimated in separate computer runs, is to place item parameters on a common scale. The item calibration models used in this study were (1) the one-parameter logistic (Rasch) model, (2) the three-parameter logistic model, and (3) an approximate three-parameter logistic model based on groups divided into fifths. The BICAL computer program (Wright & Mead, Note 2) was used to estimate item parameters for the one-parameter model; and LOGIST, for the three-parameter model. Quantile (Marco et al., Note 1), a modified version of LOGIST, was used to estimate item parameters for the approximate three-parameter model.

BICAL was used to calibrate the items from the 6 verbal and 6 mathematical data sets identified in Figure 2. A separate application of BICAL was made for each data set. BICAL estimates item parameters on a separate scale, defined by the distribution of $\underline{b}$'s, for each item calibration.

A reasonable alternative to the one-parameter logistic model is the three-parameter logistic model. This model has two potential advantages over the one-parameter logistic model discussed previously. First, for multiple-choice items a lower asymptote greater than zero is a more reasonable assumption than a lower asymptote of zero. Second, the LOGIST program takes omitted and not reached items into account, whereas the BICAL program

considers omitted and not reached items as incorrect responses. LOGIST was used to estimate item parameters for the three-parameter model.

The third item calibration method used data on examinees grouped into fifths to approximate LOGIST results. Considerable cost savings could result if the abilities for each examinee did not have to be estimated. This model was designed to calibrate items on the basis of item analysis information routinely produced at Educational Testing Service.

A new computer program Quantile was developed by modifying LOGIST to accept grouped data. The Quantile program estimates the item parameters for a test using item responses of groups of examinees instead of responses of individual examinees. The examinees are divided into groups before the program is applied. For each item the input to the program is the number of examinees in each group who answered the item correctly plus a fraction (1/(no. of response options)) of the number who omitted the item, the number of examinees in each group who reached the item, the total number of examinees who answered the item correctly, and the total number of examinees who omitted the item. All of the examinees in a group are treated as having the same ability. This ability is estimated using maximum likelihood in the same manner as individual abilities are estimated by LOGIST. The options available and the output are identical to the LOGIST options and output.

Score Equating

Once appropriate item parameter estimates were available, score equating could be accomplished. Score equating was used in this study to provide a criterion to assess the practical consequences of using the three item cali-

bration methods in conjunction with the score-conversion method of item-parameter transformation. IRT true score equating was used in each instance. This method was also used in the SAT scale stability study. Lord (1980) discussed this kind of equating in his recent book. Here only a brief summary of the method is given.

The first step in the score equating is to compute the true number right score from the item parameters that have been placed on a common scale. The true number right score is a function of the ability level θ and the item parameters a, b, and c. For the one-parameter logistic model $a = 1/1.702$ and $c = 0$ for every item. (The division by 1.702 is necessary to make the results of BICAL consistent with the results of LOGIST.) Let R stand for the number right true score. Then $R(\theta) = P_g(\theta)$, where

$$P_g(\theta) = c_g + (1 - c_g)/(1 + \exp(1.702\, a_g\, (-b_g))).$$

The true formula score is obtained by assuming that everyone answered all items, so that

$$FS = R - (N - R)/(k_g - 1),$$

where N is the number of test items and k_g is the number of response options for the item g. In the case of SAT-mathematical, which contained both four- and five-choice items two correction terms were used - one for four-choice items and one for five-choice items.

True formula scores on two tests are said to be equated if they are functions of the same θ. To obtain the raw-to-scaled score conversions for

raw scores, it is necessary only to determine θ for each FS on the test form and to find the corresponding FS on the other form. In practice, of course, estimates of the item parameters rather than the unknown true parameters are used for calculating the true scores.

Equating Form V4 to itself was accomplished in one step once the item parameters were transformed. The "final" scaled scores resulting from this equating could then be compared to the initial scaled scores. Small differences would suggest that the score-equating method of item parameter transformation is a satisfactory alternative to the use of common-item linkages.

Results and Discussion

Comparisons of Item Parameter Transformations and Associated Item Parameters

Existing score-equating parameters for converting raw scores to scaled scores were used to compute the transformations for scaling item parameter estimates. This procedure was used with the one-parameter logistic item parameters from BICAL and with the three-parameter logistic item parameters from Quantile and LOGIST to see how well this type of item parameter transformation compared with other methods of placing item parameter estimates on a common scale. It is possible that the variation in the common-item sections traditionally used for item calibration, for example, adds error to both the item-parameter estimation and transformation processes. Linking item-parameter estimates by use of existing score equating results, if satisfactory, would greatly simplify the problem of creating large pools of

items with parameters on a common scale. The transformations based on score-conversion information are shown in Table 3, along with the item parameter transformations previously derived on common items (equating tests) using BICAL and LOGIST.

Table 3 shows that the final transformations for Form V4 are very close to the identify transformation in all six situations in which score-conversion information was used to derive item-parameter transformation. The intermediate and final transformations were similar to those derived from common items.

Table 4 shows two comparisons: one is the comparison of the final (new) and initial (old) parameters and item characteristic curves for a given method, and the second, the comparison of the final parameters and item characteristic curves with the parameters of the initial LOGIST item calibration. The root mean squared error for a given item characteristic curve was computed using the ability distribution for the sample on which the initial estimates were based. It is equal to

$$[(\text{Mean Difference})^2 + (\text{SD of Difference})^2]^{1/2} ,$$

where the difference is the difference between the item characteristic curves at a given ability level. The root mean squared error for the row labeled "ICC" is the mean of the root mean squared errors across all items. The root mean squared errors for a's, b's, or c's is equal to the expression given above, but the difference is the difference in the item-parameter estimates. The ability distribution was not used for this computation.

Insert Tables 3 and 4 about here

The LOGIST (modified) item parameter estimates had a slightly smaller root mean squared error than the Quantile estimates, as might have been expected, in comparison with the initial LOGIST estimates. In the case of SAT-verbal, the LOGIST (modified) estimates were even closer to the initial estimates than the final LOGIST estimates LOGIST (Random) were. The BICAL results were on a different scale and thus could not be correlated with but otherwise compared with the initial LOGIST estimates.

The within-model comparisons also showed a small amount of error, and the results for LOGIST (modified), BICAL (modified), and Quantile demonstrate that the within-model error was smaller than the between-model error. (See the values of the mean difference, root mean squared error, and correlations.)

To summarize, the within-model and between model comparisons show that the transformed item parameter estimates based on score-conversion information agreed closely with initial estimates.

Comparisons of Score Equating Results

The results of the equatings using the item-parameter transformations based on score equating information, along with comparative results from the Petersen et al. (1983) study are shown in Table 5 and Figures 3, 4, and 5. The codes used in Figures 3, 4, and 5 are as follows:

CRIT: Criterion -- initial scaled score;

BICAL: BICAL calibrations with common-item transformations;

CONCUR: Three-parameter logistic -- LOGIST concurrent calibrations;

LOGIST(MOD): Three-parameter logistic -- LOGIST calibrations with score-conversion transformations;

BICAL(MOD): One-parameter logistic -- BICAL calibrations with score-conversion transformations;

QUANTILE: Three-parameter logistic -- Quantile calibrations based on fifths with score-conversion transformations.

Insert Table 5 and Figures 3, 4, and 5 about here

The concurrent equating model was evaluated in the SAT scale stability study (Petersen et al., 1983) and, beginning in January 1982, is being used operationally to equate SAT scores. In concurrent IRT equating item-parameter transformation is unnecessary as a separate step. Items from a new form, an old form, and a common anchor test are calibrated together using LOGIST, which produces item parameter estimates on a common scale. Then new form scores are equated to old form scores on the basis of these item parameter estimates. For the next equating the new form becomes the old form. Items from this form, another new form, and another common anchor test are calibrated together, and the scores on the total tests are equated. This kind of sequential "pairwise" equating was judged the most adequate of the three-parameter logistic IRT equating models represented in the SAT scale stability study. The results for concurrent equating referenced in this report are taken from that study.

Table 5 summarizes equating results in terms of several discrepancy indices. It gives the means and standard deviations for the scaled scores resulting from the various equatings and for the criterion scores. Ideally, the mean and standard deviation for a particular model would correspond exactly to the mean and standard deviation of the criterion scores.

The table also gives the weighted mean squared difference and its two components, the mean difference and the standard deviation of the difference. For each raw score x on Form V4 there are final scaled scores resulting from the various equatings and an initial scaled score. The smaller the differences between the final score for a particular equating model and the initial score, the more accurate the equating results are. To compute the weighted mean squared differences the values of x were weighted according to their actual occurrence in some reference group. The weighted mean squared difference is equal to the variance of the differences plus the mean differences squared. The values in Table 5 were computed from point-by-point conversions, summing over verbal raw scores 1 to 80 and mathematical raw scores -8 to 55, the ranges over which extrapolation was necessary, using the corresponding frequencies for the total group who took Form V4 when it was first administered in December 1973. The results are directly comparable with those in the SAT scale stability study (Petersen et al., 1983).

Figure 5 depicts the weighted mean squared difference in terms of its two components. The curved lines in the figure represent four levels of weighted mean squared error: 25, 100, 225, and 400. A particular point on

a line is equal to standard deviation of the difference squared plus the mean difference squared. The equating models are represented by numbers in the case of verbal equatings and by letters in the case of mathematical equatings.

Table 5 indicates that in the case of SAT-verbal the mean squared error decreased from 83.35 to 1.69 for the one-parameter model [BICAL vs. BICAL (modified)] and from 178.98 to 6.99 for the three-parameter model [concurrent LOGIST (characteristic-curve transformation) vs. LOGIST (modified)]. Large decreases in the mean squared error for SAT-mathematical are also evident. Surprisingly, the Quantile equating results had a smaller mean squared error than LOGIST (modified), even though the Quantile item-parameter estimates were inferior to the LOGIST estimates (See Table 4). One can infer than the use of item parameter transformations derived from score equating information was very effective, resulting in reducing mean squared error.

The differences in mean squared errors among the models utilizing score-equating transformations is slight and of little practical importance, for the scores almost always rounds to the same 200-800 scaled score, which is reported with a third digit of 0.

Conclusion

The use of item parameter transformations based on score conversion information needs further study before they can be applied operationally in testing programs. In this study the chain was short, and half of the common-item linkages used in the study were the very same ones that had been

used in score equating. This probably created a situation in which the conversion parameters were more consistent with item calibration results than would be expected if items were calibrated for test forms more widely separated in the genealogical chart. One needs to find out how well these transformations work when items from a variety of old forms, particularly those linked together by long equating chains, are calibrated. In conclusion, item parameter transformations based on score-conversion information look promising, but need further testing under more realistic conditions.

Reference Notes

1. Marco, G. L., Wingersky, M. S., & Douglass, J. B. An evaluation of three approximate item response theory models for equating test scores. Unpublished manuscript, Educational Testing Service, 1984.

2. Wright, B. D., & Mead, R. J. BICAL: Calibrating items with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1976.

References

Lord, F. M. A broad-range tailored test of verbal ability. Research Bulletin 75-5. Princeton, N.J.: Educational Testing Service, 1975.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N. J.: Erlbaum, 1980.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Petersen, N. S., Cook, L. L., & Stocking, M. L. Item response theory versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 1983, 8, 137-156.

Ree, J. M. Estimating item characteristic curves. Applied Psychological Measurement, 1979, 3, 371-385.

Stocking, M. L., & Lord, F. M. Developing a common metric in item response theory. Applied Psychological Measurement, 1983, 7, 201-210.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.) Application of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.

Wright, B. D., & Stone, M. Best test design: Rasch measurement. Chicago, Ill.: MESA, 1979.

Footnote

[1]The data for the study were the scored item tapes used by Petersen, Cook, and Stocking (1983) and derived from records of the College Board Admissions Testing Program, administered by ETS. The author wishes to thank Edwin O. Blew, Karen Carroll, and Marilyn S. Wingersky for processing and analyzing the data used in the study.

Table 1

Summary Statistics for SAT-Verbal and SAT-Mathematical Equating Samples

		Verbal				Mathematical			
				Scaled Score[a]				Scaled Score[a]	
Form	Admin. Date	Equating Test	N	Mean	SD	Equating Test	N	Mean	SD
V4	12/73	fe	2665	438	114	ff	2628	457	115
X2	4/75	fe	2686	437	106	ff	2629	476	111
X2	4/75	fm	2562	432	106	fn	2527	471	111
Y3	6/76	fm	2578	426	112	fn	2553	465	113
Y3	1/78	fw	2549	405	109	fx	2455	443	117
B3	5/79	fw	2700	433	108	fx	2633	479	114
B3	5/79	fk	2665	429	104	f1	2596	476	111
Y2	4/76	fk	2879	432	108	f1	2815	469	115
Y2	4/76	fu	2774	428	105	fv	2721	472	115
Z5	12/77	fu	2853	414	108	fv	2774	447	114
Z5	12/77	et	2814	417	110	eu	2739	444	113
V4	12/73	et	2670	436	113	eu	2673	455	115

[a]Scaled score statistics are linear transformations of raw score statistics and deviate slightly from reported score statistics in those cases where curvilinear transformations were used operationally.

Table 2

Scaling Parameters for SAT-Verbal and SAT-Mathematical Forms[a]

SAT Form	Verbal		Mathematical	
	A	B	A	B
V4	6.9931	193.2421	8.8839	272.5925
X2	6.9304	193.0270	8.6393	265.4098
Y3	6.8813	189.1449	8.4892	260.3593
B3	6.8315	183.8779	8.5734	267.4198
Y2	7.2588	184.5344	8.5533	269.7831
Z5	6.9440	200.4501	8.4740	273.1202

[a]The 200 to 800 College Board scaled score (S) is determined by the formula $S = AX + B$, where X is the raw score (number right corrected for guessing).

Table 3

Transformations for Placing Item Difficulties from Different Item Calibration Models on a Common Scale

Scale Relationship	One-Parameter		Three-Parameter		
				Score-Conversion	
	Common-Item (BICAL)	Score-Conversion (BICAL (Modified))	Common-Item (LOGIST (Char.-Curve))[a]	LOGIST (Modified)	Quantile (Fifths)
SAT-Verbal					
X2 - V4	$\underline{b}$ - .090	.877$\underline{b}$ - .175	.898$\underline{b}$ - .040	.931$\underline{b}$ - .038	.938$\underline{b}$ - .049
Y3 - V4	$\underline{b}$ - .174	.861$\underline{b}$ - .188	.950$\underline{b}$ - .296	.995$\underline{b}$ - .312	1.013$\underline{b}$ - .326
B3 - V4	$\underline{b}$ - .214	.849$\underline{b}$ - .177	.839$\underline{b}$ - .073	.897$\underline{b}$ - .065	.916$\underline{b}$ - .068
Y2 - V4	$\underline{b}$ - .130	.906$\underline{b}$ - .166	.847$\underline{b}$ - .087	.908$\underline{b}$ - .082	.925$\underline{b}$ - .081
Z5 - V4	$\underline{b}$ - .130	.876$\underline{b}$ - .065	.903$\underline{b}$ - .231	.997$\underline{b}$ - .184	1.011$\underline{b}$ - .214
V4 - V4	$\underline{b}$ - .122	1.014$\underline{b}$ - .046	.898$\underline{b}$ - .092	.979$\underline{b}$ + .002	.994$\underline{b}$ - .018
SAT-Mathematical					
X2 - V4	$\underline{b}$ - .131	.996$\underline{b}$ - .159	.948$\underline{b}$ + .041	.993$\underline{b}$ + .066	.986$\underline{b}$ + .064
Y3 - V4	$\underline{b}$ - .226	.950$\underline{b}$ - .205	1.068$\underline{b}$ - .211	1.109$\underline{b}$ - .215	1.127$\underline{b}$ - .243
B3 - V4	$\underline{b}$ - .182	1.021$\underline{b}$ - .074	.942$\underline{b}$ + .064	1.005$\underline{b}$ +.106	.985$\underline{b}$ + .109
Y2 - V4	$\underline{b}$ - .189	.956$\underline{b}$ - .113	.959$\underline{b}$ + .031	1.048$\underline{b}$ + .058	1.039$\underline{b}$ + .058
Z5 - V4	$\underline{b}$ - .116	.971$\underline{b}$ - .101	.999$\underline{b}$ - .259	1.051$\underline{b}$ - .224	1.097$\underline{b}$ - .233
V4 - V4	$\underline{b}$ - .014	1.010$\underline{b}$ - .013	.998$\underline{b}$ - .083	1.010$\underline{b}$ - .026	1.010$\underline{b}$ - .016

[a]Transformation procedure used in the Petersen, Cook, and Stocking (1983) SAT scale drift study.

Table 4

Comparisons of New (Transformed) Item Parameter Estimates with Old (Untransformed) Estimates and Criterion Estimates from LOGIST[a,b]

	Parameter	Mean Crit.	S.D. Crit.	Mean Diff.	RMSE[c]	r[d]	Mean Old	S.D. Old	Mean Diff.	RMSE[c]	r[d]
SAT-Verbal											
LOGIST (Random)	a	.782	.272	-.042	.122	.91	[Same as Criterion Comparisons]				
	b	.304	1.306	.046	.193	.99					
	c	.153	.060	.003	.044	.70(.92)					
	ICC				.022						
LOGIST (Modified)	a	[Same as above]		-.026	.118	.91	[Same as Criterion Comparisons]				
	b			.041	.181	.99					
	c			.003	.044	.70(.92)					
	ICC				.021						
BICAL (Modified)[e,f]	a	[Same as above]				-	.588	.000	-.008	.008	-
	b					.97	-.064	1.437	.006	.087	.99(8)
	c					-	.000	.000	.000	.000	-
	ICC										
Quantile (Fifths)	a	[Same as above]		-.020	.161	.83	.748	.287	.014	.157	.85
	b			-.020	.235	.99	.277	1.355	.007	.157	.99
	c			.003	.054	.57(.88)	.148	.053	.008	.037	.78(.94)
	ICC				.025					.021	
SAT-Mathematical											
LOGIST (Random)	a	.968	.333	-.039	.156	.89	[Same as Criterion Comparisons]				
	b	.350	1.184	-.004	.159	.99					
	c	.130	.042	-.020	.040	.69(.85)					
	ICC				.026						
LOGIST (Modified)	a	[Same as above]		-.048	.158	.89	[Same as Criterion Comparisons]				
	b			-.028	.166	.99					
	c			-.020	.040	.69(.86)					
	ICC				.026						
BICAL (Modified)[e,f]	a	[Same as above]				-	.588	.000	-.006	.006	
	b					.97	.031	1.531	.005	.123	.99(7)
	c					-	.000	.000	.000	.000	
	ICC										
Quantile (Fifths)	a	[Same as above]		-.003	.173	.89	.957	.379	.007	.143	.93
	b			-.023	.171	.99	.311	1.153	.017	.140	.99
	c			.003	.031	.70(.90)	.127	.035	.006	.027	.73(.88)
	ICC				.029					.024	

[a] New item-parameter estimates were derived from data on the sample that took Form V4-et; old estimates and criterion estimates, from data on the sample that took Form V4-fe.

[b] New and old item-parameter estimates were derived by use of procedures for the specified model; criterion estimates were derived by use of LOGIST.

[c] Root Mean Squared Error = $\{[\text{Mean(Difference)}]^2 + [\text{SD(Difference)}]^2\}^{\frac{1}{2}}$, based on statistics that took Form V4-fe.

[d] The correlation coefficients in parentheses are the correlations of c's for those items on which c's were estimated and sufficient data existed to yield a reasonably good estimate of the c's ($b-2/a > -2$).

[e] For BICAL only, the correlation of the b's is given for the comparison with the criterion estimates; the different metrics used for the item-parameter estimates precluded meaningful comparisons of the other statistics. Also, the correlations of the new and old a's and the new and old c's were undefined because of the lack of variance.

[f] For BICAL, root mean squared errors were not calculated; ability estimates were not available in the dataset.

BEST COPY AVAILABLE

Table 5

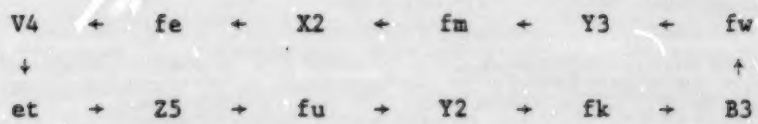
Information and Summary Discrepancy Indices[a] for Selected Item Response Theory Equating Models

Index	Initial Scale (Criterion)	LOGIST (Concurrent)	Common-Item Transformation		Score-Conversion Transformation		
			LOGIST (Characteristic-Curve)	BICAL	LOGIST (Modified)	BICAL (Modified)	Quantile (Fifths)
SAT-Verbal							
Scaled Score:							
Mean	435.37	445.73	446.34	444.45	434.50	434.33	434.93
Standard Deviation	109.09	112.89	116.47	109.60	106.61	108.33	108.58
Mean Squared Error[b]		125.15	178.98	83.35	6.99	1.69	5.75
Mean Difference		10.36	10.97	9.08	-.87	-1.04	-.44
S.D. of Difference		4.23	7.66	.96	2.50	.79	2.36
SAT-Mathematical							
Scaled Score:							
Mean	468.05	471.61	474.95	467.40	468.49	467.39	467.81
Standard Deviation	113.25	115.50	114.82	113.32	112.87	112.36	113.43
Mean Squared Error[b]		23.27	61.08	.46	7.28	1.31	1.61
Mean Difference		3.55	6.90	-.66	.44	-.66	-.24
S.D. of Difference		3.26	3.67	.18	2.66	.93	1.25

[a] Computed for SAT-verbal raw scores 1 through 80 and for SAT-mathematical raw scores -8 through 55.

[b] $(\text{Mean Difference})^2 + (\text{SD of Difference})^2$.

Verbal Chain



Mathematical Chain

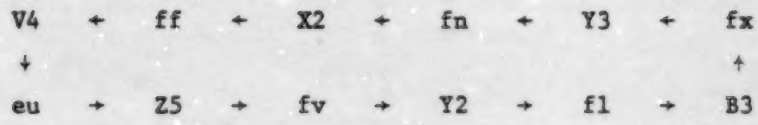


Figure 1. Verbal and mathematical item calibration chains.

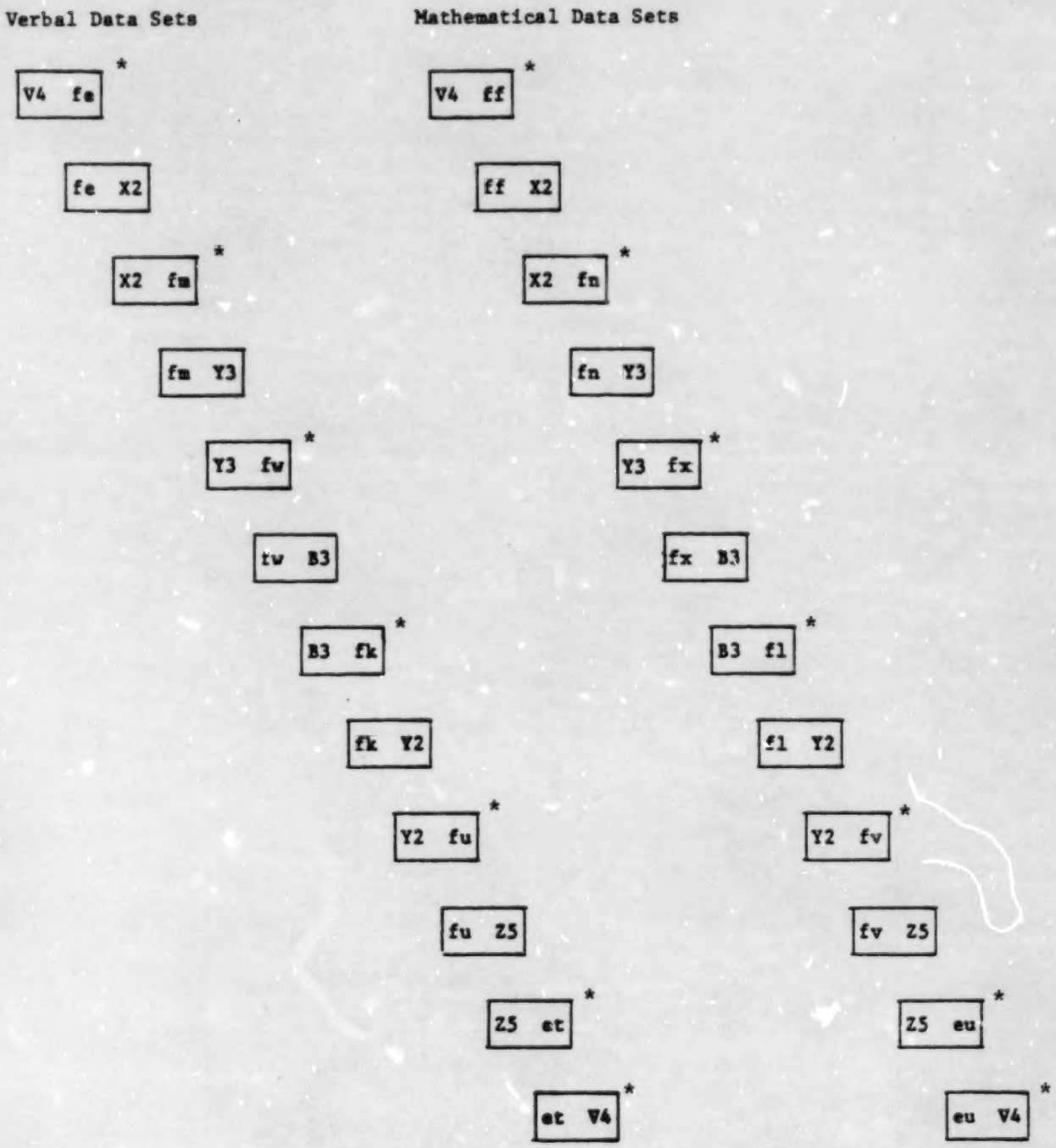
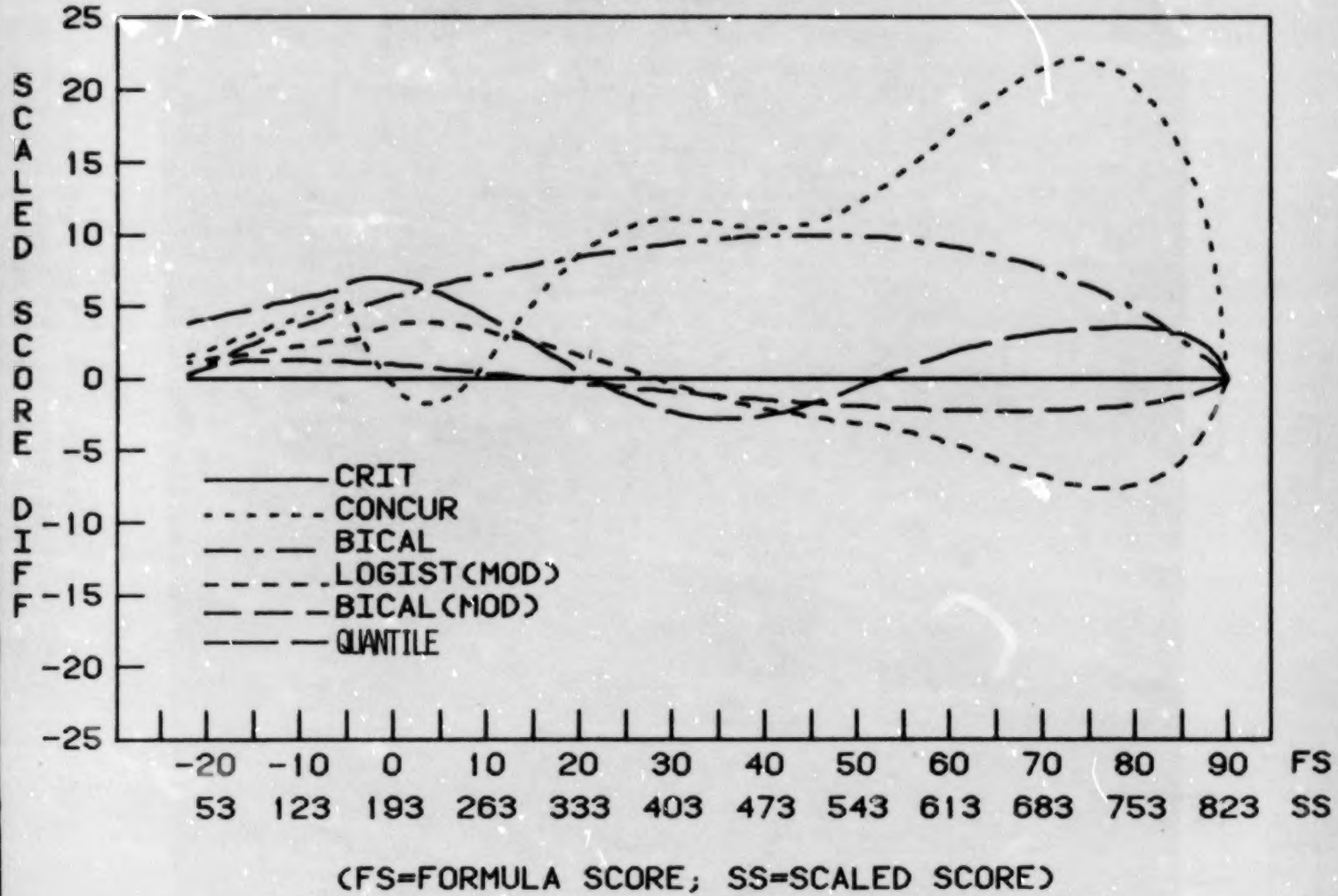


Figure 2. Verbal and mathematical data sets. Each box represents a sample of approximately 2,670 cases. The asterisks identify the data sets used in deriving item parameter transformations from score-conversion information.

FIGURE 3

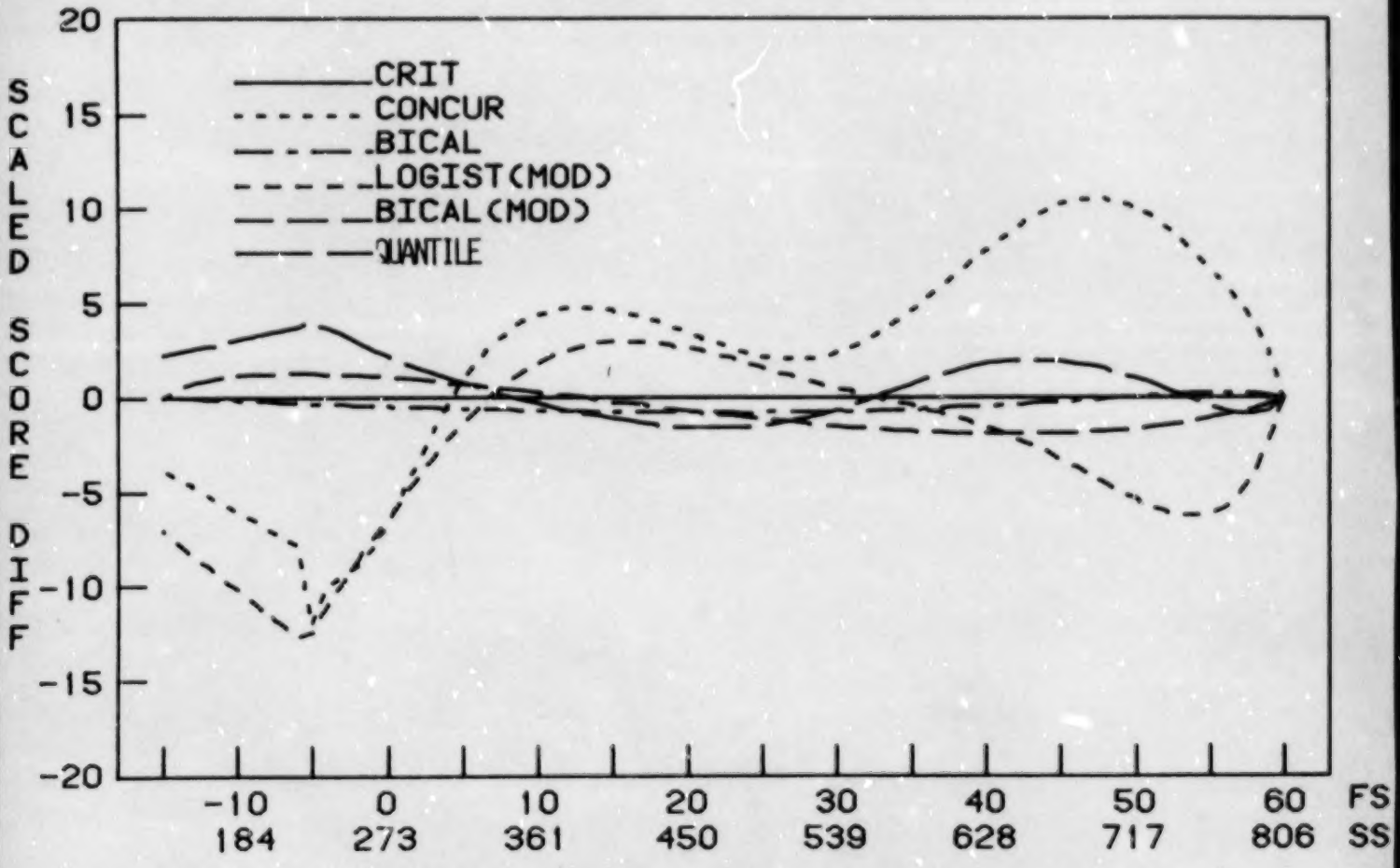
SAT-VERBAL EQUATING LINES
(MODEL MINUS CRITERION)



(FS=FORMULA SCORE; SS=SCALED SCORE)

FIGURE 4

SAT-MATHEMATICAL EQUATING LINES
(MODEL MINUS CRITERION)



(FS=FORMULA SCORE; SS=SCALED SCORE)

FIGURE 5

MEAN SQUARED ERROR (M.S.E.)
FOR APPROXIMATE IRT EQUATING MODELS